**Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Agnes implementation on lung capacity data

Submitted by

Ravina.M 1MS15CS102

Vaibhav Srivastava 1MS15CS135

Namrata Balachandra 1MS15CS155

*Bachelor of Engineering in Computer Science & Engineering*

Under the guidance of

Sowmya B J

Assistant Professor

CSE department

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

www.msrit.edu, **20177**

**Introduction**

Our dataset brings out the information about smoking in the children and teenagers. It depicts the lung capacity of various age groups of people depending on their gender. The average total **lung capacity** of an adult human male is about 6 litres of air. Tidal breathing is normal, resting breathing; the tidal volume is the volume of air that is inhaled or exhaled in only a single such breath. Dataset compares lung capacity with smoking and several other factors.

In [data mining](https://en.wikipedia.org/wiki/Data_mining) and [statistics](https://en.wikipedia.org/wiki/Statistics), **hierarchical clustering**  is a method of [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) which seeks to build a [hierarchy](https://en.wikipedia.org/wiki/Hierarchy) of clusters. Strategies for hierarchical clustering generally fall into two types:

* **Agglomerative**: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
* **Divisive**: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

**Datasets Description**

**Source of Dataset**- www.Kaggle.com

**Attribute Description**

Attributes in our dataset Lung Capacity are Age, LungCap, Height, Gender, Smoke.

* Age- Is the Age of that Particular Person, In our data set we have ages between 4-19.
* LungCap- Which is Lung Capacity of that particular person.
* Height-Height of that particular person.
* Gender-Gender of that particular person (Male or Female).
* Smoke-If that particular person smokes (yes or no).

**Data Set Size and Number of Tuples**

* Data Set -Lung Capacity
* Lung Capacity Data Set size in terms of Bytes-14.2 KB (14,595 bytes)
* Number of Tuples in Lung Capacity Data Set- 655 Tuples

**Inference:**

* Lung capacity increases with increase in height.
* Male’s lung capacity is higher compared to female lung capacity.
* Lung capacity in children is less compared to a teenager.
* Smokers have less lung capacity compared to non- smokers.
* Above 9 years children smoke.

**Algorithm Description**

**Agnes:**

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc.

The hierarchy within the final cluster has the following properties:

* Clusters generated in early stages are nested in those generated in later stages.
* Clusters with different sizes in the tree can be valuable for discovery.

**Algorithm:**

This algorithm  works by  grouping  the data one by one on the basis of the  nearest distance measure of all the pair wise distance between the data point. This works according to bottom up approach.

Step 1: Consider X={x1,x2,……,xn} to be a set of datapoints.

Step 2: Initially we are in the level 0 and sequence number=0.

Step 3: Find the distance between all points.

Step 4: Cluster the points according to the least distance.

Step 5: Increment the sequence number and find similarity between the smaller clusters and cluster similar clusters into bigger clusters.

Step 6:Continue the same until you get the bigger cluster which forms the root.

**Function:**

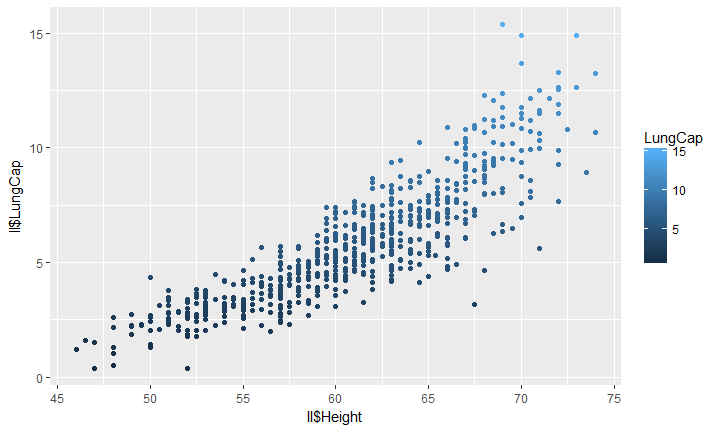
*agnes(x, diss = False, metric = "euclidean")*

**Arguments:**

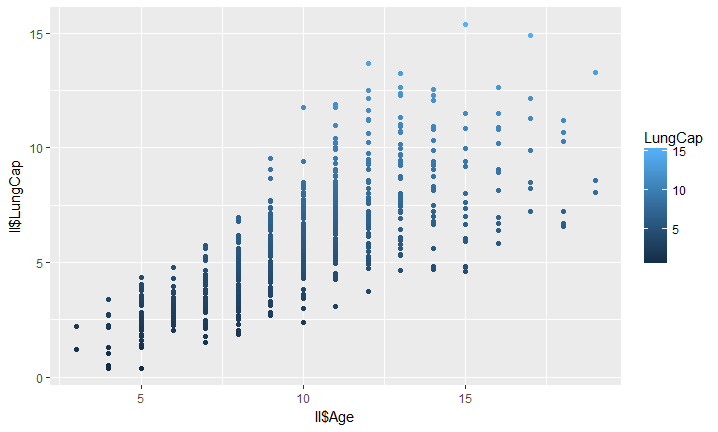
|  |  |
| --- | --- |
| X: | data matrix or data frame, or dissimilarity matrix, depending on the value of the diss argument.  In case of a matrix or data frame, each row corresponds to an observation, and each column corresponds to a variable. All variables must be numeric. Missing values (NAs) are allowed.  In case of a dissimilarity matrix, x is typically the output of [daisy](https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/daisy.html) or [dist](https://stat.ethz.ch/R-manual/R-devel/library/stats/html/dist.html). Also a vector with length n\*(n-1)/2 is allowed (where n is the number of observations), and will be interpreted in the same way as the output of the above-mentioned functions. Missing values (NAs) are not allowed. |
| diss | logical flag: if TRUE (default for dist or dissimilarity objects), then x is assumed to be a dissimilarity matrix. If FALSE, then x is treated as a matrix of observations by variables. |
| metric | character string specifying the metric to be used for calculating dissimilarities between observations. The currently available options are "euclidean" and "manhattan". Euclidean distances are root sum-of-squares of differences, and manhattan distances are the sum of absolute differences. If x is already a dissimilarity matrix, then this argument will be ignored.  **Snapshot of the code** |

**Result Snapshot and Description**

1)By plotting a graph with Lung Capacity and Height we can Infer that as Height increases lung Capacity increases as shown in the graph below

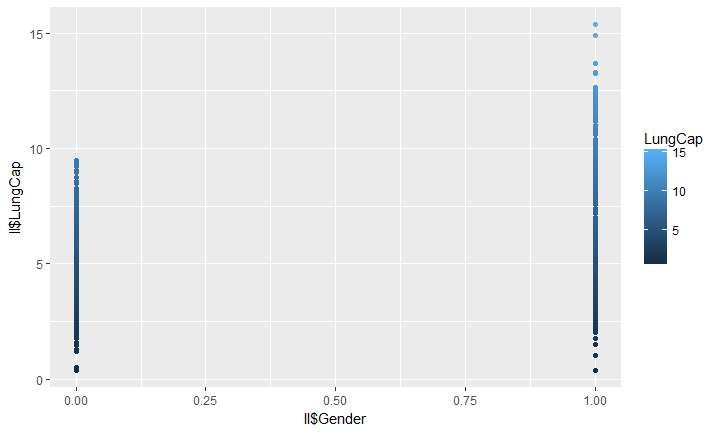
Therefore Lung Capacity is directly proportional to Height( Lung Capacity α Height) with increase in Height lung Capacity increases.

2)By plotting a graph with Lung Capacity and age we can infer that between age 4-7 children have less lung capacity and as they grow at the age of 18 lung capacity is stable and increases.



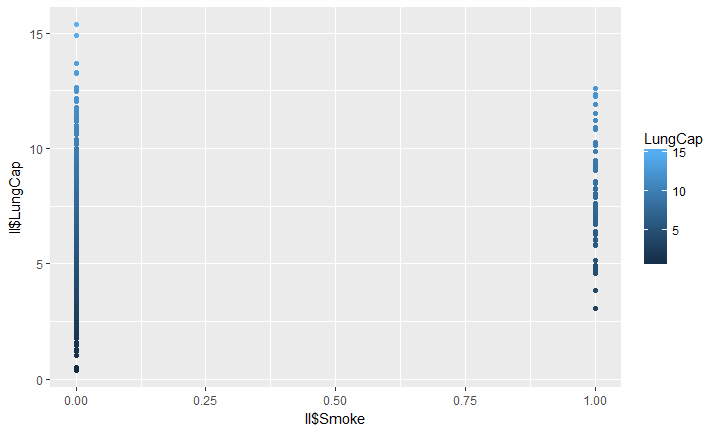
Therefore as age increases to the age of 18 years lung Capacity is more Compared to the children between the age of 4-7 years.

3)By plotting a graph with Lung capacity and age we can infer that the lung capacity of male is much higher compared to females.



Where male is 1 and female is 0 hence from the graph above we can infer that male have higher lung capacity than females.

4)by plotting a graph between lung capacity and smoke we can infer that people who smoke have less lung Capacity than people who don’t smoke

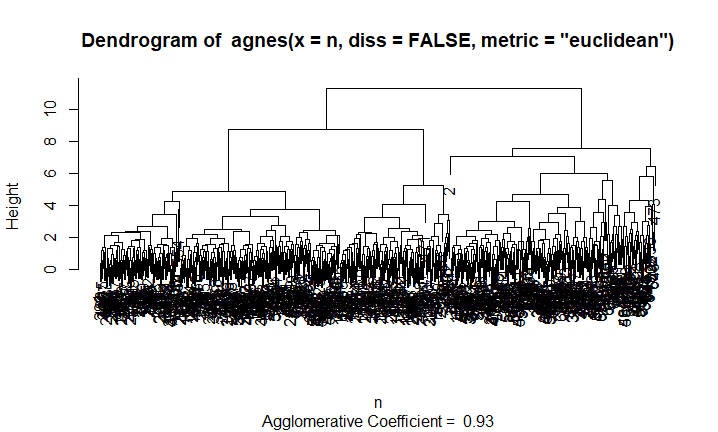


Where ,people who don’t smoke is 0 and who do smoke is 1.therfore smokers have less lung Capacity than non smokers.

5)by plotting a graph between smoke and age we can infer that children with age ranging from 4-9 do not smoke and some children above 9 years smoke .



Agnes dendrogram graph



**Implant**

Step 0: Include library ggplot and cluster.

Step 1:Importing the dataset(LungCapacity) into R by using: setwd("C:/Users/namra\_000/Desktop/Agnes") and rr= read.csv('LungCapData2.csv') where the dataset is stored in the variable rr.

Step 2: Changed categorical values to numerical values(Gender(Female=0 and Male=1) and Smoke(Yes=1 and No=0)) by using rr$Gender<-ifelse(rr$Gender=="male",1,0) and rr$Smoke<-ifelse(rr$Smoke=="yes",1,0)

Step 3: Converting the dataset to dataframe by using the statement:

n<-as.data.frame(rr)

Step 4: Agnes function:

n1<-agnes(n,diss=FALSE,metric="euclidean")

Step 5:Finding the correlation between different combination of attributes by using

cor(rr$atrribute1,rr$attribute2)

Step 6: Plot( n1) and the result will be a dendogram where agnes function is stored in n1 variable.

Step 7: We plot a graph using the function ggplot where the data can be inferred:

ggplot(rr,aes(rr$Attribute1,rr$Attribute2,color=LungCap))+geom\_point()

After converting the data set attribute values to numeric

Age LungCap Height Gender Smoke

1 9 3.124 57.0 0 0

2 8 3.172 67.5 0 0

3 7 3.160 54.5 0 0

4 9 2.674 53.0 1 0

5 9 3.685 57.0 1 0

6 8 5.008 61.0 0 0

7 6 3.757 58.0 0 0

8 6 2.245 56.0 0 0

9 8 3.961 58.5 0 0

10 9 3.826 60.0 0 0

11 6 2.806 53.0 0 0

12 8 3.205 54.0 1 0

13 8 4.579 58.5 0 0

14 8 4.354 60.5 1 0

15 8 4.774 58.0 1 0

16 7 3.796 53.0 1 0

17 5 2.416 50.0 1 0

18 6 3.634 53.0 0 0

19 9 5.056 59.0 1 0

20 9 5.812 61.5 1 0

21 5 2.200 49.0 0 0

22 5 1.768 52.5 0 0

23 4 0.517 48.0 0 0

24 7 5.734 62.5 1 0

25 9 6.964 65.0 0 0

26 3 2.212 51.5 1 0

27 9 5.044 60.0 1 0

28 5 3.265 52.0 1 0

29 8 6.940 60.0 0 0

30 9 4.300 60.0 0 0

31 5 1.846 49.0 0 0

Code

library(cluster)

library(ggplot2)

setwd("C:/Users/namra\_000/Desktop/Agnes")

rr= read.csv('LungCapData2.csv')

#importing the dataset

rr$Gender<-ifelse(rr$Gender=="male",1,0)

rr

rr$Smoke<-ifelse(rr$Smoke=="yes",1,0)

rr

n<-as.data.frame(rr)

n1<-agnes(n,diss=FALSE,metric="euclidean")

cor(rr$Height,rr$LungCap)

cor(rr$Age,rr$LungCap)

cor(rr$Gender,rr$LungCap)

cor(rr$Smoke,rr$LungCap)

cor(rr$Height,rr$LungCap)

cor(rr$Smoke,rr$LungCap)

cor(rr$LungCap,rr$Smoke)

cor(rr$Gender,rr$LungCap)

cor(rr$Age,rr$Smoke)

ggplot(rr,aes(rr$Height,rr$LungCap,color=LungCap))+geom\_point()

ggplot(rr,aes(rr$Age,rr$LungCap,color=LungCap))+geom\_point()

ggplot(rr,aes(rr$Smoker,rr$LungCap,color=LungCap))+geom\_point()

ggplot(rr,aes(rr$Gender,rr$LungCap,color=LungCap))+geom\_point()

ggplot(rr,aes(rr$LungCap,rr$Age,color=LungCap))+geom\_point()

ggplot(rr,aes(rr$Smoke,rr$Age,color=Age))+geom\_point()

levels(rr$Age)

plot(n1)

Social Impact

Lung Capacity refers to the amount of air in lungs after taking taking the deepest breath possible. Earlier symptoms of reduced lung capacity includes shorteness of breath, decreased stamina ,reduced endurance and frequent and respiratory infections. A normal adult has vital capacity between age 4 and 7 as age increases the lung capacity increases. Body size,especially height has huge huge impact on lung capacity,as height increases lung capacity increases so people with good height has good lung capacity and when we compare smokers and non-smokers,non-smokers have good lung capacity.Male have higher lung capacity when compared to female.

Children should be given good nutritious diet and make them involve in activities such as basketball,swimming etc,to help them attain good height which internally increases their lung capacity.We need to spread awareness among children about the harmful effects of smoking ,so that they can prevent smoking and hence they will have a good health with good lung capacity